网络舆情观点主题识别研究*

李　真　丁晟春　王　楠
(南京理工大学信息管理系　南京 210094)

摘要: 【目的】识别网络舆情中的观点主题。【方法】通过舆情信息内容、用户关系、用户行为三个方面的 4 个维度(时间维、用户维、内容维、观点维)的关联, 构建微博舆情观点主题识别模型。【结果】提出包括舆情网络构建、观点主题抽取及聚类、"用户–所属观点主题" 2-模网络构建、观点主题演化分析 4 部分的网络舆情观点主题识别方法体系, 实验结果证明该方法体系可有效识别网络舆情中的观点主题。【局限】用户属性对观点主题识别的影响有待进一步考虑。【结论】基于社会网络视角, 利用 LDA 主题模型, 可多方面、多维度地识别网络舆情观点主题。

关键词: 网络舆情　社会网络　LDA 模型　主题识别　观点主题
分类号: TP391　G350

1　引　言

自媒体平台在给人们提供共享、交流新方式的同时, 也使得由网络引起、放大或主导的社会舆情事件频发。由于互联网具有信息发布及时、传播速度快、影响范围广等特性, 导致舆情事件一旦在互联网上爆发将呈不可逆转的趋势。此外, 网络的开放性和隐蔽性为网民提供了观点表达的场所, 观点是人们对某个事物或事件所产生的带有情感倾向性的看法或态度。面对海量的网络舆情观点信息, 政府和企业要想及时做好网络舆情引导工作, 就必须快速把握网络舆情参与主体当下所持有的主要观点。本文将网民主体在舆情事件中所处立场, 或者所形成的对舆情事件/问题的主要看法称为网络舆情观点主题, 而从大规模网络舆情信息中获取观点主题, 并进行展示的一系列技术方法就称作观点主题识别。

现有研究多是根据网络舆情发展结果进行滞后性的动因分析、回溯分析、演化分析, 处理方式处于被动的问题解决状态, 不能满足政府和企业应急管理中实时监测舆情动态的现实要求。此外, 用户在微博、贴吧等自媒体网络平台上发布的信息具有数据类型多样化、文本内容碎片化与不完备等特性, 使得传统的舆情事件研究方法不能满足现有网络舆情分析的需要。因此网络舆情的研究方法需要得到进一步的创新, 在文本内容处理的基础上, 重视用户行为、用户关系等社会化特征数据, 多维度地挖掘网络舆情中的主题。

基于此, 本文基于社会网络视角, 利用 LDA 主题模型, 引入时间变量, 提出一种动态识别网络舆情参与主体所持观点主题变化情况的模型, 以期为政府和企业的网络舆情监测和引导提供理论支持, 满足关键舆情事前跟踪和事中实时发现的需求。

2　相关研究

目前有关网络舆情主题识别的研究呈现出较为迅速的递增趋势, 由于网络舆情传播途径的多样性, 研究者除针对不同类型的网络舆情信息开展主题识别研究外, 还基于主题特征的差异性, 针对不同类型的主题展开研究。

2.1　网络舆情主题识别研究现状

国外关于网络舆情主题识别的研究起步较早, 采用的主题识别研究方法也更为多样。Wu 等利用 TF-IDF 算法、Text-Rank 算法, 提取微博关键词, 标注

通讯作者: 丁晟春, ORCID: 0000-0002-4269-021X, E-mail: todingding@163.com。
*本文系国家社会科学基金项目"基于社会网络分析的网络舆情主题发现研究"(项目编号: 15BTQ063)的研究成果之一。

用户兴趣爱好，以挖掘用户兴趣及关注热点[1]。Narang 等则是将 TF-IDF 算法与文本聚类以及 WordNet 局部相似性检测相结合，发现围绕主题的社交对话[2]。Kim 等使用 Twitter 数据进行实验，发现词频比率能够恰当地检测社交热点话题或突发新闻[3]。Nguyen 等提出一个社交软件平台，用于从 Twitter 等类似的社交网络服务的信息扩散模式中检测出有意义的事件，实现热门话题的发现[4]。Guo 等利用 FrequentPattern 流挖掘算法实现 Twitter 热点主题的检测[5]。

国内对舆情主题识别的研究主要基于聚类思想，或通过改进 LDA 模型，利用单一维度的文本信息实现对网络舆情主题的挖掘。如叶川等利用 LDA 主题模型进行热点评论的分类推断及主题特征挖掘，实现对微博热门评论的主题标签推断[6]。唐晓波等针对文本聚类和 LDA 主题模型的互补特征，提出一种两者结合的微博主题检索模型[7]。伍万坤等对标准 LDA 模型进行改进，提出一种挖掘电商微博热点话题的 EM-LDA 综合模型[8]。部分文本聚类方法虽然兼顾了文本内容的结构信息和语义信息，但很难充分表达语义信息。而利用 LDA 主题模型实现网络舆情主题识别的研究多侧重于对 LDA 主题模型的改进，仍是单一维度的主题识别。此外，还有研究者将本体论和语义计算的相关技术引入到网络舆情事件的主题识别中[9]，同时还融入影响力计算、句法依存、社会网络分析等改进方法[10-14]，进一步完善网络舆情主题识别研究。

微博具有信息发布便捷、快速、实时等特点，逐渐成为网络舆情爆发的主阵地，以微博为主要研究对象的舆情主题识别研究占据了半壁江山。从微博主题识别的类型上看，已有研究多是对微博社区主题、微博热点主题的挖掘，少量涉及到对观点主题、潜在主题的研究。其研究方法从初始的简单聚类逐步演化到通过 LDA 主题模型结合词汇、句子、时间、情感等特征辅助实现对微博主题的检测，这些研究方法多是从主题词途径来识别舆情主题，没有综合考虑网络舆情中用户的社会信息，以及用户行为对于舆情传播演化的影响。

2.2 网络舆情观点主题识别研究现状

观点主题识别是指从大规模的观点性评论信息中获取主题，并进行展示的一系列技术方法的总称，旨在从海量的评论信息中迅速获得用户对某一舆情事件或问题的主要看法和态度。对舆情观点信息的挖掘研究多倾向于观点的抽取与识别。从观点抽取的结果来看，可将现有研究大致分为三类。

(1) 抽取舆情观点所指对象，如周杰等提出一种领域无关的由内到外的观点主题识别算法[15]，其中观点主题是指观点所指对象，仅能指明网民大众所评论的焦点，并不能直观表示出网民大众对该讨论焦点所持的主要看法或态度。

(2) 按舆情观点的情感倾向进行观点的分类。丁晟春等结合心理学与自然语言处理技术，将微博情绪分为喜、怒、哀、恶、惧 5 大类，利用情感特征、句式特征及句间特征对微博情绪进行表示，借助 SVM 模型形成微博情绪 5 类分类模型，实现微博情绪的多类分类[16]。这种观点挖掘仅从宏观上把握网民主体的情感极性，并不能体现观点的具体描述。

(3) 舆情观点词或观点句的识别及描述。陈晓美等运用多文档文摘技术和以句子为单位的 LDA 主题模型方法，获得每个主题具有代表性的观点言论，揭示网络舆情主要观点[17]。姚兆旭等利用 LDA 模型和改进的 TF-IDF 算法构建主题特征词向量，基于相似度计算自动抽取主题词汇链，在此基础上，引入情感词典，实现主题观点词的抽取[18]。无论是将语法语义相结合的观点识别方法，还是将情感极性与主题信息相结合的方法，实质上都是基于内容的观点识别，其忽略了社交网络平台中用户行为、用户属性等社会化特征数据，不能满足现有微博舆情观点主题识别的需要。

基于以上研究现状，本文将在文本内容处理的基础上，引入社会化网络分析途径，将主题模型与社会网络分析相结合，实现从海量评论信息中获取微博舆情参与主体的主流观点，为政府与企业的微博舆情引导工作提供有力的理论支撑。

3 网络舆情观点主题识别建模

本文选取新浪微博为研究平台，通过舆情信息内容、用户关系、用户行为三个方面的 4 个维度(时间维、用户维、内容维、观点维)的关联，构建网络舆情观点主题识别模型。

3.1 维度设计

(1) 时间维。已有研究往往选取舆情事件生命周

期作为时间粒度，即对舆情事件的潜伏期、成长期、爆发期和衰退期等各个阶段进行分析，以实现对舆情传播过程的动态监测。本文所研究的观点主题——网民对某一事件所持观点的变化往往发生在更短的时间粒度内，即在舆情生命周期的某一阶段内观点已发生多次变化。因此本文以“天”为时间粒度，研究每天观点主题的变化情况。

(2) 用户维。用户是网络舆情产生的主体，网络舆情事件正是由于用户在互联网上表达对该事件的认知、态度和意见，并进行传播而形成的。本文对用户维度的研究侧重于用户的行为及用户间评论、点赞等关系，具体包括三类:

①发布行为。当微博用户想要及时分享所见所闻、发表个人观点时会产生发布行为，用户的发布行为促使个人观点的产生。

②评论行为。当微博用户对原始微博所述内容感兴趣或持有个人看法时，会发生评论行为，评论行为也会促使观点的产生。

③点赞行为。当微博用户对正在浏览的原始微博或评论表示赞同时会产生点赞行为，是行为成本最低的观点表达行为。

(3) 内容维。本文所指“内容”表示的是用户发表的以微博或微博评论为载体的带有情感倾向性的文字内容，其中，将用户直接发布的微博内容称为“原始微博”，将评论内容称为“评论微博”。这些带有用户情感倾向性的内容里隐含了用户的观点，是观点的具体阐释，即内容一定包含了用户的某一种或某几种观点。本文允许用户发表多条内容，但假设每条内容仅包含一种观点。

(4) 观点维。观点通常是指用户对某一事件或事物所持有的看法或情感倾向，并不是舆情事件的基本要素，而是基于内容总结、提取得到的，即观点是从用户发表的内容中高度概括、总结出来的。

图 1 展示了时间、用户、内容及观点 4 个维度间的关系。随着时间维度的变化，舆情事件参与主体、发表内容、所持观点及网民大众整体观点倾向都会不断发生变化。有的用户会出于兴趣等原因参与舆情事件整个生命周期过程(如用户 A)，也有用户只在某一阶段参与了对该事件的讨论(如用户 B)。用户在参与事件讨论的不同阶段会发表一条或几条不同的内容，可能是发表了不同内容但表达了同一观点(如用户 C 发表的内容 C2、C3)，也有可能是发表了不同内容且表达了多种观点(如用户 D 发表的内容 D2、D3)。有的用户虽发表了多条内容，但内容所含观点始终不变(如用户 C 始终持有观点 3，直至用户 C 退出此事件的讨论)，也有用户发表内容的所属观点会随时间的推移而发生变化(如用户 A 在 T1 时期发表的内容 A1 属于观点 1，在 T2 时期发表的内容 A2 属于观点 4)。

本文不研究具体某一用户的观点变化情况，而是研究参与事件讨论的网民整体所持观点主题的变化情况。

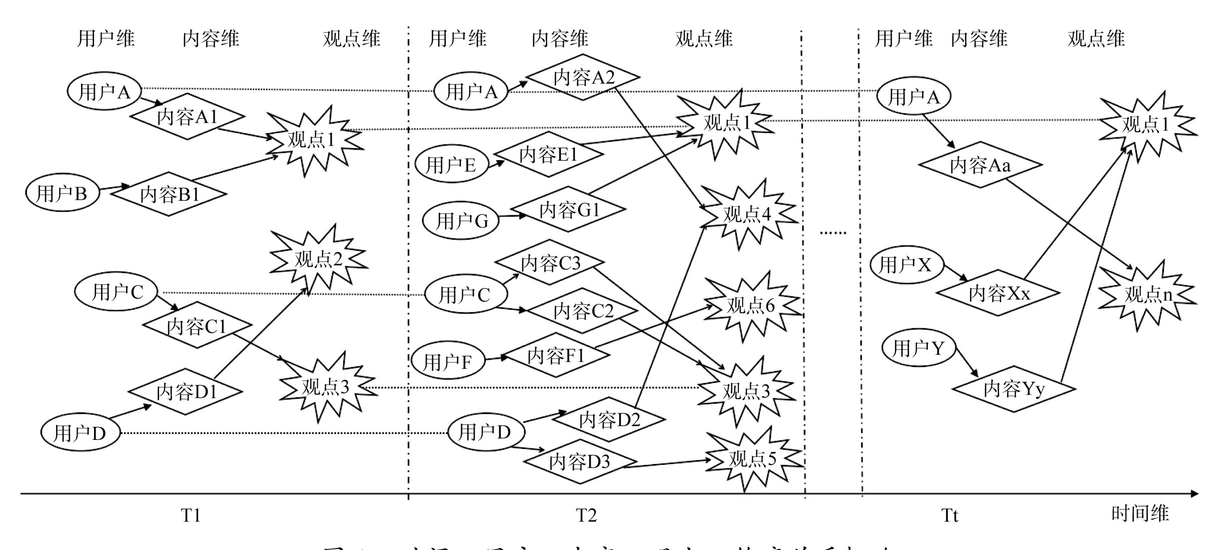


图 1　时间、用户、内容、观点四维度关系概略

3.2　总体框架

本文网络舆情观点主题识别框架包括舆情网络构建、观点主题抽取及聚类、“用户–所属观点主题”2-模网络构建、观点主题演化分析 4 部分，如图 2 所示。

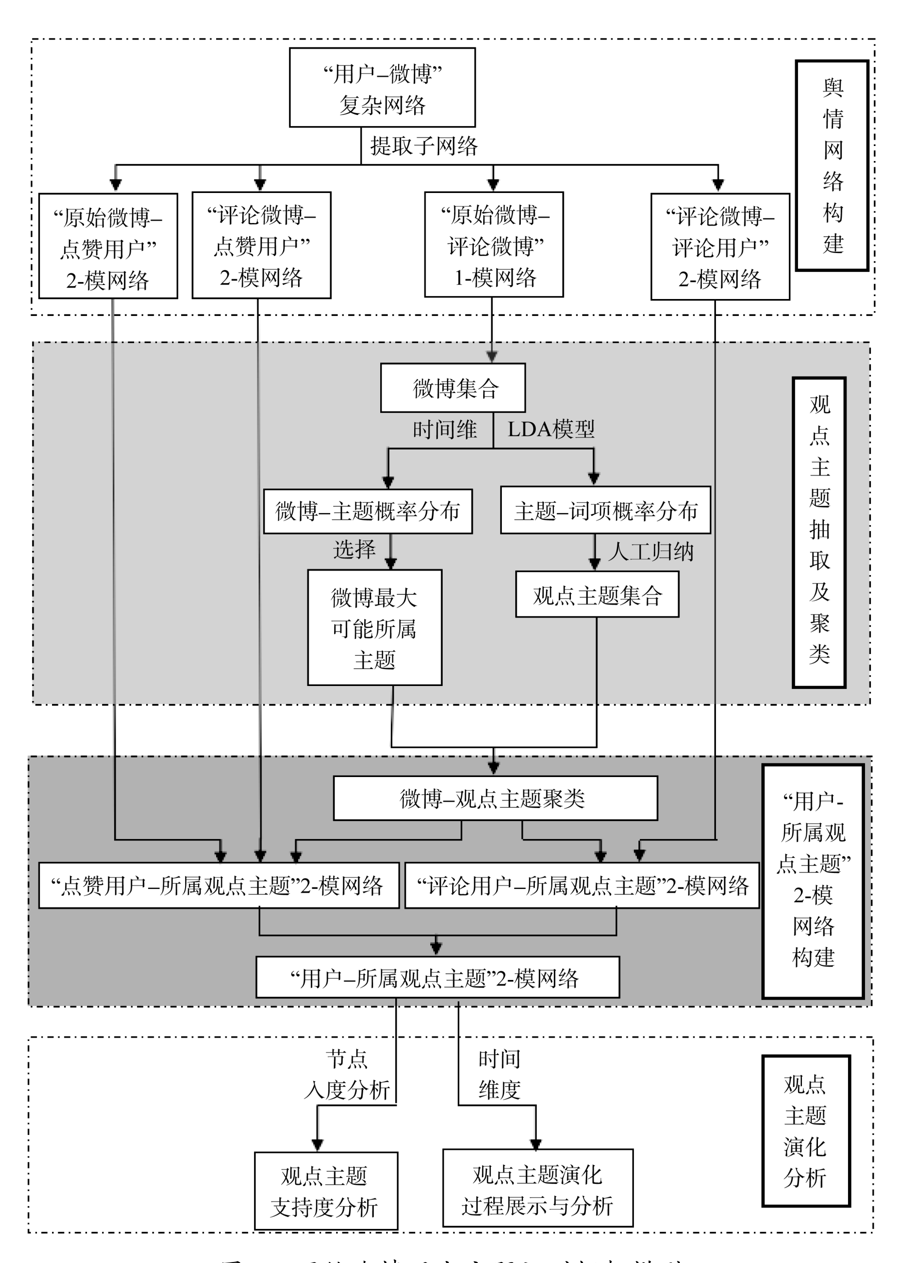


图 2　网络舆情观点主题识别框架模型

(1) 基于用户维和内容维构建“用户–微博”复杂网络，并从该复杂网络中提取 4 个子网络，其中“原始微博–评论微博”1-模网络节点构成了微博集合，即该模型要处理的文本集合。根据 3.1 节中对时间维度的分析，以“天”为时间粒度，将待处理的文本集合离散到相应的时间窗口，依次处理各时间窗口的文本，并进行下一步的主题抽取。

(2) 利用 LDA 模型得到该微博集合的“微博–主题”概率分布和“主题–词项”概率分布。依据“主题–词项”概率分布，将各主题下的词项进行人工归纳得到各主题所代表的观点，即本文所研究的观点主题，形成观点主题集合，完成对微博舆情观点主题的抽取；

同时，根据“微博–主题”概率分布矩阵，利用 LDA 直接聚类，实现对微博按主题聚类。

(3) 在上述微博聚类结果的基础上，结合网络舆情构建中提取出的其他三个子网络中微博与用户的对应关系，构建“用户–所属观点主题”2-模网络。

(4) 基于“用户–所属观点主题”2-模网络进行舆情观点主题演化分析，该演化分析包括两部分：一是通过对“用户–所属观点主题”2-模网络中节点入度的分析得到观点主题支持度排名，对每天的观点主题支持度变化情况进行分析与说明；二是对舆情事件观点主题“产生–发展–衰退”的演化过程进行展示与分析。

3.3 “用户–微博”复杂网络构建

“用户–微博”复杂网络中的节点包括微博、用户两大类，其中微博节点包括原始微博和评论微博两种；用户类节点包括评论用户及点赞用户，其中点赞用户又分为点赞原始微博的用户和点赞评论微博的用户。节点关系主要涉及回复、发表和点赞三种关系。节点与节点间的关系如图 3 所示。

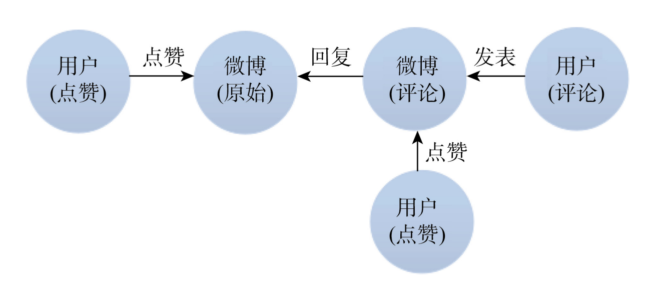


图 3 “用户–微博”复杂网络模型

(1) “原始微博–评论微博”1-模网络

为了分析原始微博的观点和评论微博的观点，将“用户–微博”复杂网络的用户节点剔除，提取回复关系，转换成“原始微博–评论微博”1-模网络，如图 4 所示。该网络将原始微博和评论微博视为同一种节点，即微博节点，以它们之间的回复关系为连线，连线的方向代表了回复的方向。同时，一条评论微博仅代表对原始微博的一条回复，所以线值为 1。该网络由一个个不连通的“星型”子网络构成，以“原始微博”为子网络的中心节点，外围的“评论微博”必须通过中心节点才能建立联系，即评论微博与评论微博之间因共同回复了同一原始微博才建立联系。因此，该网络的网络密度低，网络规模取决于外围节点个数，在一定程度上表征了原始微博的热度。

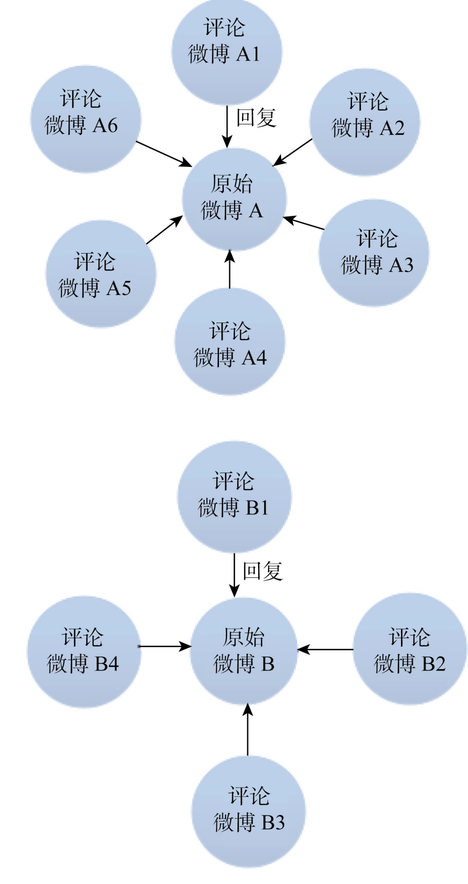


图 4 “原始微博–评论微博”1-模网络模型

(2) “原始微博–点赞用户”2-模网络

提取图 3 中原始微博和点赞用户两类节点，连线代表节点间的点赞关系，连线方向代表点赞方向。由于用户对同一条微博只能发生一次点赞行为，因此连线线值为 1。图 5 为“原始微博–点赞用户”2-模网络结构，该网络呈“网状”结构，用户间因点赞了同一微博而建立联系，同时，不同原始微博之间因被同一用户点赞而建立联系。该网络的网络规模取决于点赞用户节点个数，不仅在一定程度上表征了原始微博的热度，也表达了用户对原始微博的赞同强度。

(3) “评论微博–点赞用户”2-模网络

现实生活中，对评论和原始微博点赞的用户可能存在交集，但由于难以获取对评论进行点赞的用户的具体信息，因此本文假设对评论点赞的用户群和对原始微博点赞的用户群不存在交集。提取图 3 中的评论微博和点赞用户两类节点，以其之间的点赞关系为连线，连线的方向代表点赞的方向。由于用户对同一条

评论只能发生一次点赞行为, 因此连线线值为 1。图 6 为该 2-模网络模型结构, 与图 5 所示网络不同点在于: 该网络会因原始微博的不同而形成不同的子群, 子群内的网络密度高于子群间的网络密度, 即对同一条原始微博的不同评论进行点赞的用户的重合度更高。评论微博节点数量和点赞用户节点数量在一定程度上表征了原始微博热度, 点赞用户节点数量体现了评论微博被赞同的强度。

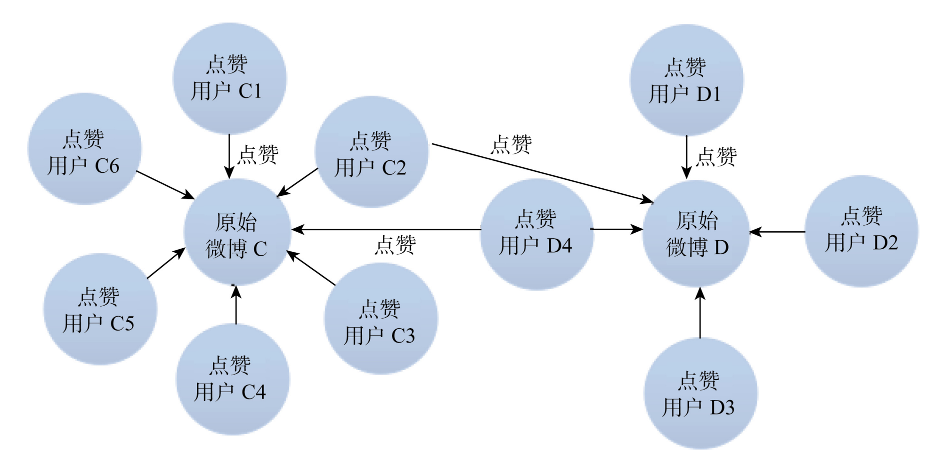


图 5 “原始微博–点赞用户”2-模网络模型

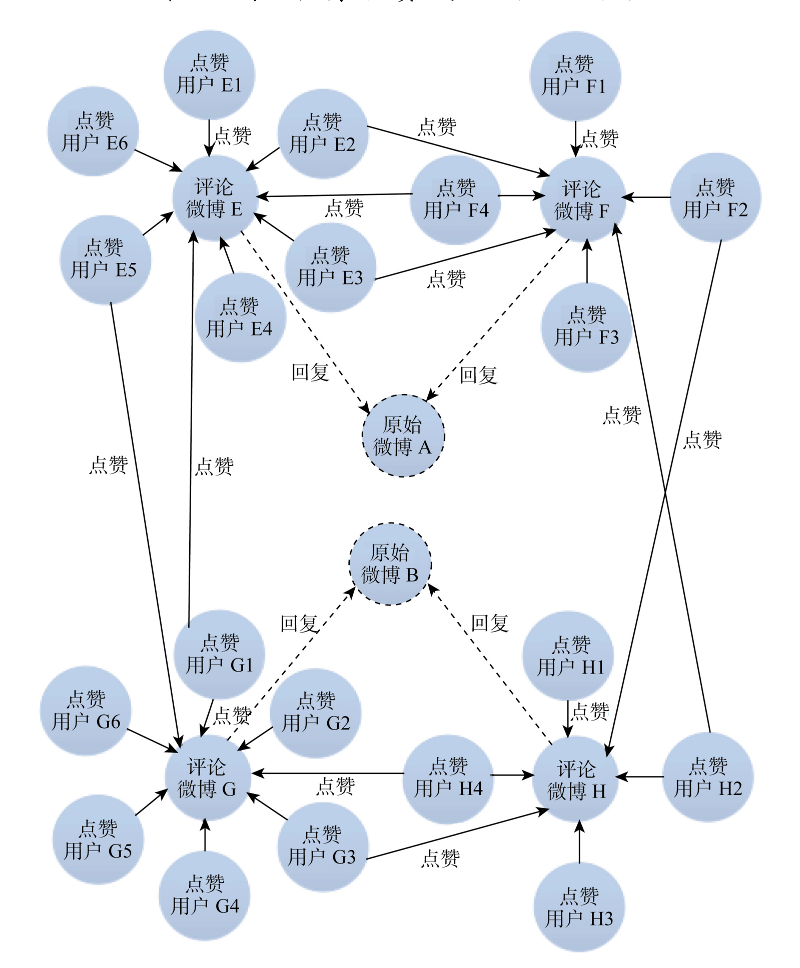


图 6 “评论微博–点赞用户”2-模网络模型

(4) “评论微博–评论用户”2-模网络

提取图 3 中的评论微博和评论用户两类节点，以它们之间的发表关系作为连线，连线方向代表发表关系的方向。虽然用户可以就同一个原始微博发表多次评论，但由于本文将每条评论视作一条单独的微博，所以评论用户与评论微博间也是一一对应关系，即连线的线值为 1。图 7 为“评论微博–评论用户”2-模网络模型，同图 4 所示网络类似，由一个个不连通的“星型”子网络构成，以“评论用户”为子网络的中心节点，外围的“评论微博”必须通过中心节点才能建立联系，即评论微博与评论微博之间因由同一用户发表才建立了联系。评论是一种行为成本较高的用户行为，因此网络规模，即外围评论微博节点个数，在很大程度上表征了用户的活跃程度。

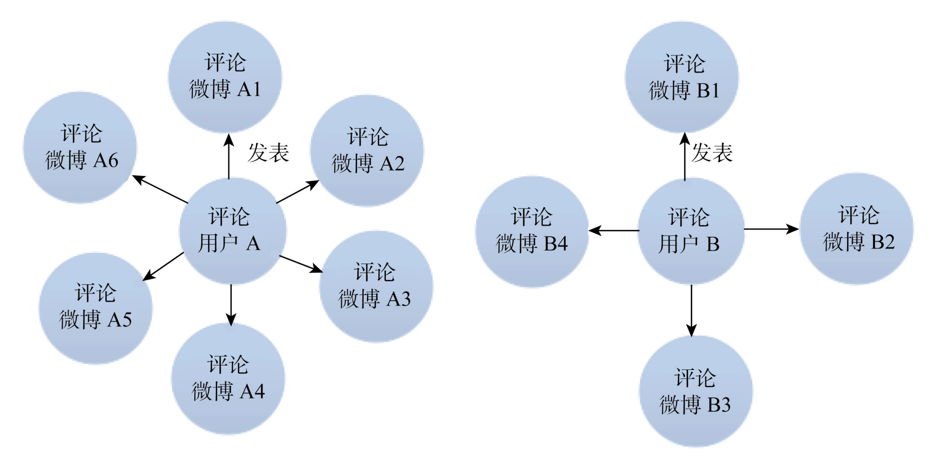


图 7 “评论微博–评论用户”2-模网络模型

3.4 微博观点主题抽取及聚类

LDA 模型认为文档是由主题按一定概率组成的，而每个主题又是若干词项的概率分布。利用 LDA 模型进行主题的抽取是上述文档生成过程的逆过程，通过 LDA 模型可以得到“主题–词项”概率分布和“文档–主题”概率分布。由于本文选取的研究对象是带有观点倾向性的评论微博和原始微博，因此对 LDA 模型抽取出来的微博主题词项进行归纳，组织成句，即视为微博观点主题。微博观点主题抽取过程如图 8 所示。

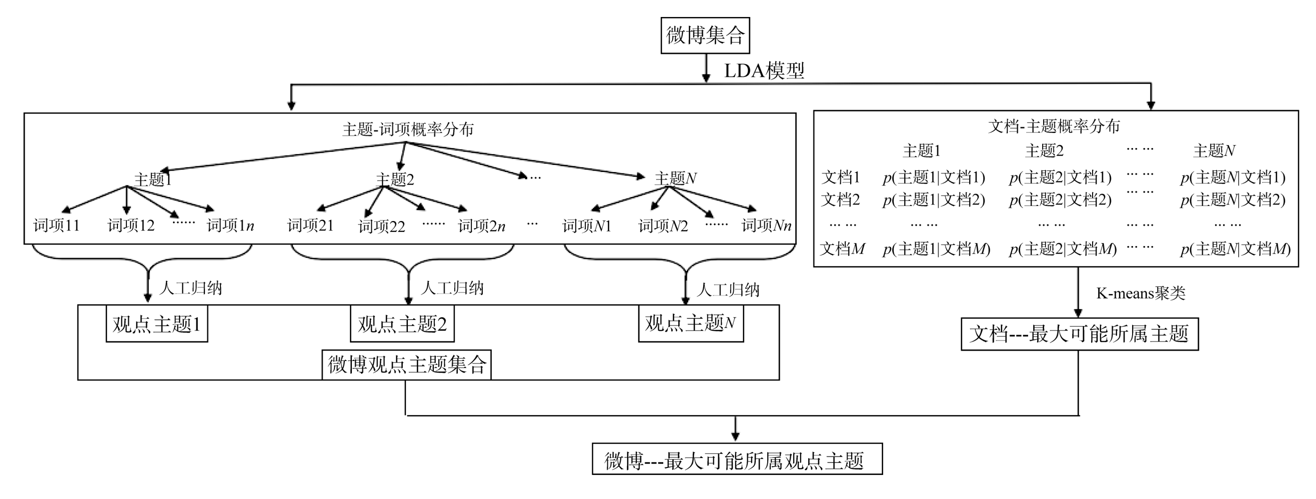


图 8 微博观点主题抽取模型

其中，利用LDA主题模型进行微博舆情观点抽取的过程主要包括文本预处理、文本建模、主题特征词提取、主题词合并归纳等步骤，最终得到主题–词项概率分布，同时得到文档–主题分布，表现形式如下。

$$
\begin{matrix}
p(z_1|d_1) & p(z_2|d_1) & \cdots & p(z_k|d_1) \\
p(z_1|d_2) & p(z_2|d_2) & \cdots & p(z_k|d_2) \\
\vdots & \vdots & \ddots & \vdots \\
p(z_1|d_M) & p(z_2|d_M) & \cdots & p(z_k|d_M)
\end{matrix}
$$

利用该文档–主题分布得到每条微博文本的最大可能所属观点主题，实现微博聚类。对于第 n 个文本，即矩阵的第 n 行，若有 $\overline{t} = \arg\max_{1\leqslant t\leqslant k} p(z_t|d_n)$，则将第 n 个文本归入主题 t 中。

3.5 "用户–所属观点主题"2-模网络构建

基于 3.4 节微博聚类结果及 3.3 节用户与微博的对应关系，构建"用户–所属观点主题"2-模网络，如图 9 所示。该网络模型说明一个观点主题的受支持程度取决于表达该观点主题的评论用户数与赞同该观点的点赞用户数之和。

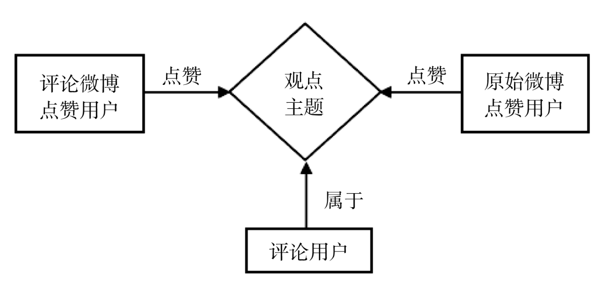


图 9 "用户–所属观点"2-模网络模型

3.6 基于"用户–所属观点主题"网络的观点支持度分析

基于内容维度提取出微博舆情观点主题后，为了得到网民最支持的观点，还需要加入用户维度，从社会网络的视角对各观点主题的受支持程度做进一步分析。由于图 9 所示的网络模型中只涉及由用户节点指向观点主题节点的单向弧，本文选取节点入度作为观点主题支持度的测量指标。观点主题节点的入度可细分为绝对观点支持度和相对观点支持度，其中，绝对观点支持度计算如公式(1)所示。

$$s_D(o_k) = \sum_i C_{ik} + \sum_j C_{jk} = \sum_i C_{ki} + \sum_j C_{kj} \quad (1)$$

其中，o_k 表示"用户–所属观点主题"网络中的观点主题节点，$C_{\cdot k}$ 的取值范围为 $\{0,1\}$，C_{ik} 取 1 时表示点赞用户赞同该观点主题 k，C_{jk} 取 1 时表示评论用户发布表达该观点主题的相关评论。为进一步分析不同时间段(每天)的观点支持度演化情况，需要对观点支持度进行标准化处理，以做不同网络间的比较。观点主题的相对观点支持度计算如公式(2)所示。

$$S_R(o_k) = \frac{S_D(o_k)}{N_U} \quad (2)$$

其中，N_U 表示网络中的用户节点总数。

观点支持度反映了用户对该观点主题的支持程度，入度越大代表有越多的用户赞同、支持该观点，该观点主题就越有可能为主流观点，在舆论引导过程中就越应该引起政府与企业的重视。

4 实验与结果分析

以"双汇进口美国猪肉"事件为例，选取目前受众较广泛的新浪微博作为研究平台，以"双汇 猪肉"为关键词，检索并收集发布时间在 2016 年 4 月 1 日–2016 年 4 月 15 日的所有微博与评论，得到有效原始微博和评论微博共计 215 条，进行舆情观点主题识别的实验。

4.1 "双汇进口美国猪肉"事件"用户–微博"2-模网络

基于 3.3 节构建"用户–微博"2-模网络，利用 Pajek 软件对其实现可视化，效果如图 10 所示。

粉色节点表示用户类节点，蓝色节点表示微博类节点，节点大小在一定程度上表征了微博热度。由于本文选取的舆情事件规模较小，使得不同微博间的评论用户群重合度低，网络整体呈现不连通状态。

4.2 "双汇进口美国猪肉"事件舆情观点主题抽取及聚类

以"天"为时间单位，将微博离散到不同的时间窗口，选用开源的 JGibbLDA 实现每日微博主题的抽取，主题及其词项分布结果(以 2016 年 4 月 7 日为例)如图 11 所示。根据主题及各主题下最优词项的抽取结果，对词项进行合并、归纳得到观点主题，结果如表 1 所示(以 2016 年 4 月 7 日为例)。

为评价 LDA 主题抽取效果，研究对所收集的 215 条微博进行内容分析，人工总结其观点主题。通过与人工总结的主题进行比对发现，LDA 主题抽取效果较好。但利用 LDA 进行观点主题抽取主要存在两个不足。

(1) 利用 LDA 模型抽取出来的不同主题可能表达的是同一个含义。如表 1 所示的观点主题 2 与观点主题 9 都表达了"双汇进口美国猪肉没错，不是卖国贼"的意思，但 LDA 模型将其视为不同的两个主题，而人工总结时则会将两者视为同一主题。

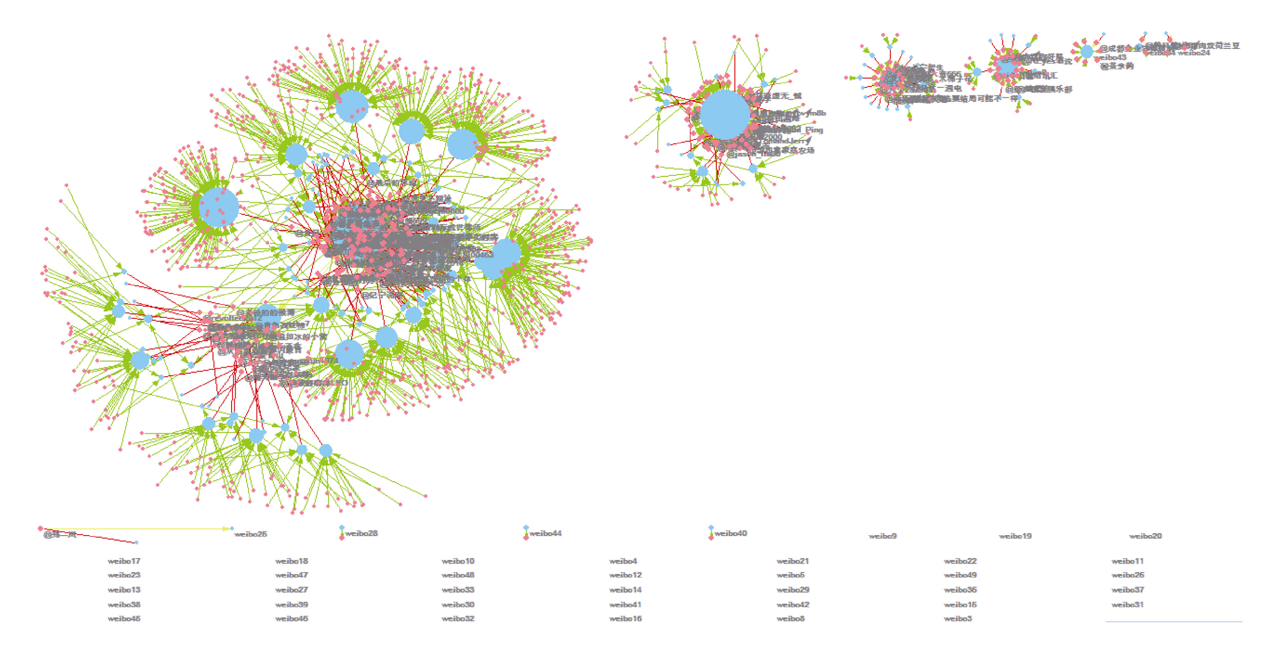


图 10 “双汇进口美国猪肉”事件“用户–微博”2-模网络可视化效果

(2) 文档预处理阶段，如“是”、“不是”这类字眼会被过滤掉，导致对LDA模型提取出来的词项进行人工归纳概括时很有可能出现总结出来的观点与实际观点恰恰相反的情况。如用户评论内容为“双汇不是卖国贼”，而LDA提取出的词项则是“双汇”、“卖国贼”，这就导致在人工归纳时，有可能将这一观点主题总结为“双汇是卖国贼”。

```
Topic 0th:
    猪肉 1.3315746753246753
    国内 0.8648538961038961
    贵 0.5280032467532467
    国外 0.22362012987012986
    埋怨 0.20738636363636365
Topic 1th:
    买 1.1991604477611941
    卖国贼 0.3782649253731343
    日本 0.3269589552238806
    香港 0.30830223880597013
    奶粉 0.30830223880597013
Topic 2th:
    不代表 0.17317541613316262
    利益 0.13796414852752883
    美国 0.13476312419974393
    政府 0.13156209987195905
    价值观 0.12836107554417414
Topic 3th:
    进口 0.40557163531114326
    农业 0.2897973950795948
    补贴 0.257235890014718
    市场 0.19934876989869757
    滞销 0.16678726483357456
Topic 4th:
    国内 0.4995378927911275
    国外 0.2869685767097967
    生活 0.23613678373382624
    企业 0.222273567467652 5
    进口 0.21765249537892792
Topic 5th:
    猪肉 0.41853932584269665
    真 0.2999375780274657
    养殖业 0.20318352059925093
    人士 0.18757802746566793
    双汇 0.13139825218476905
Topic 6th:
    高 0.3364882506527415
    物价 0.3201697127937337
    东西 0.3136422976501306
    差 0.2744778067885117 5
    造 0.2712140992167102
Topic 7th:
    双汇 0.5759361997226076
    收购 0.2604022191400832
    价格优势 0.2604022191400832
    中国 0.2395977808599168
    美帝 0.1945214979195562
Topic 8th:
    买 0.9774159663865546
    便宜 0.8723739495798318
    爱国 0.8198529411764706
    便宜货 0.3786764705882352
    人性 0.3786764705882352
Topic 9th:
    便宜 0.9365763546798029
    农产品 0.8318965517241379
    海关 0.8011083743842364
    收税 0.7826354679802955
    值得 0.7826354679802955
```

图 11 “双汇进口美国猪肉”事件评论主题最优词项提取结果(2016-4-7)

表 1 “双汇进口美国猪肉”事件观点主题(2016-4-7)

观点主题
1. 猪肉国内贵，国外便宜
2. 进口猪肉就像日本买电饭煲，香港买奶粉，不是卖国贼
3. (卖国贼说法)哗众取宠，不代表政府和社会主流价值观，不值得关注
4. 政府应该对农业进行补贴，控制市场
5. 国内物价都比国外高，愿意去国外生活
6. 双汇采用真猪肉
7. 国内物价高，东西造的质量差
8. 双汇收购是因为美帝生猪有价格优势
9. 买便宜东西是人性使然，是爱国行为
10. 国外农产品远渡重洋，经海关收税后还比国内便宜，值得深思

为进一步分析各观点主题受支持程度，需利用公式(1)实现微博与观点主题间的映射，即完成微博–观点主题聚类。表2展示了2016年4月7日的部分聚类结果。

表 2 “双汇进口美国猪肉”事件部分聚类结果(2016-4-7)

微博编号	所属观点主题编号	微博编号	所属观点主题编号
1	Topic2	115	Topic2
2	Topic1	116	Topic7
3	Topic2	117	Topic6
4	Topic9	118	Topic8
5	Topic6	119	Topic5
6	Topic10	120	Topic4

对每日微博进行人工观点分类，将得到的LDA聚类结果进行对比，结果一致用数字 1 表示，结果不一致用数字 0 表示。用对比结果一致的数量与总数之比表示聚类结果的准确率，计算得到每日微博聚类准确率，如表 3 所示。

表 3　聚类准确率(2016 年)

日期	4 月 2 日	4 月 7 日	4 月 8 日	4 月 9 日	4 月 9 日之后	平均准确率
准确率	0.66	0.53	0.71	0.47	0.6	0.56

从表 3 可以看出，用单一的 LDA 模型直接聚类方法得到的聚类结果并不十分准确，究其原因可能为：LDA 模型本身高度依赖词频；同一评论可能表达多种观点，而该聚类方式默认将评论只归为某一种观点主题。

4.3　“用户–所属观点主题” 2-模网络构建

基于 3.5 节构建的“用户–所属观点主题” 2-模网络，利用 Pajek 软件可视化，效果如图 12 所示(以 2016 年 4 月 7 日为例)。

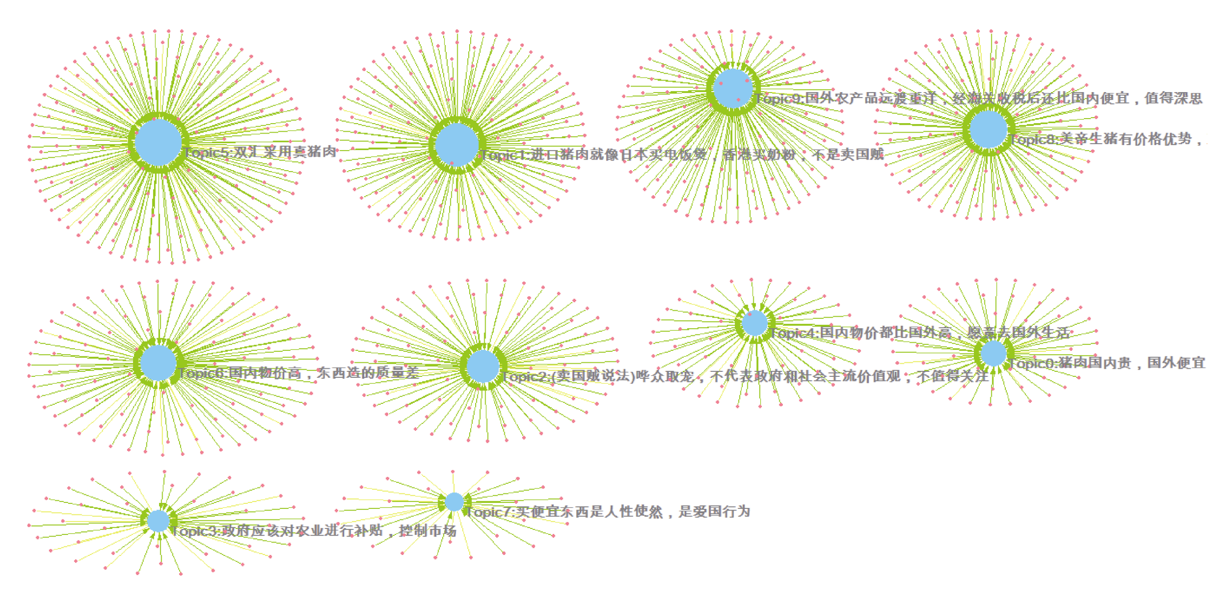


图 12　“双汇进口美国猪肉”事件“用户–所属观点主题”可视化效果(2016-4-7)

图 12 中粉色节点代表用户节点，蓝色节点代表观点主题节点，蓝色节点的大小表征了 2016 年 4 月 7 日当天观点主题的受支持度情况。由于在 3.3 节中假设点赞用户群体间不存在交集，因此该网络形成多个以各观点主题为中心的网络子群，整体呈不连通状态。对单日“用户–所属观点主题”网络的分析无法得到观点主题是如何随时间变化的，因此还需进一步的观点主题演化分析。

4.4　微博观点主题演化分析

根据 3.6 节对观点主题支持度的描述，利用节点入度求得观点主题支持度，并做归一化处理，处理结果如表 4 所示(以 2016 年 4 月 7 日为例)。可以看出，利用 LDA 模型仅基于内容维度提取出的主题排名与加入社会化数据后得到的主题排名并不一致，这说明对网络舆情的研究不能缺少对用户行为、用户关系等社会化数据的分析。在实际舆情监测中，需要加强对支持度高的观点主题的关注，尤其是当支持度高的观点为负面倾向时，更应引起政府和企业的重视，及时做好舆情引导工作，以免这些支持度高的负面观点影响网民整体情感倾向和负面情绪的二次爆发。

表 4　“双汇进口美国猪肉”事件观点主题编号及其相对支持度(2016-4-7)

观点主题编号	节点入度归一化	观点主题
6	0.19	Topic5
7	0.15	Topic1
8	0.13	Topic9
9	0.13	Topic8
10	0.11	Topic6
11	0.10	Topic2
12	0.06	Topic4
13	0.06	Topic0
14	0.04	Topic3
15	0.03	Topic7

本文以“天”为时间粒度，对各个观点主题相对支持度的变化情况进行分析。为更直观地得到各观点主题的变化情况，对其演化过程进行可视化展示，如图13所示。

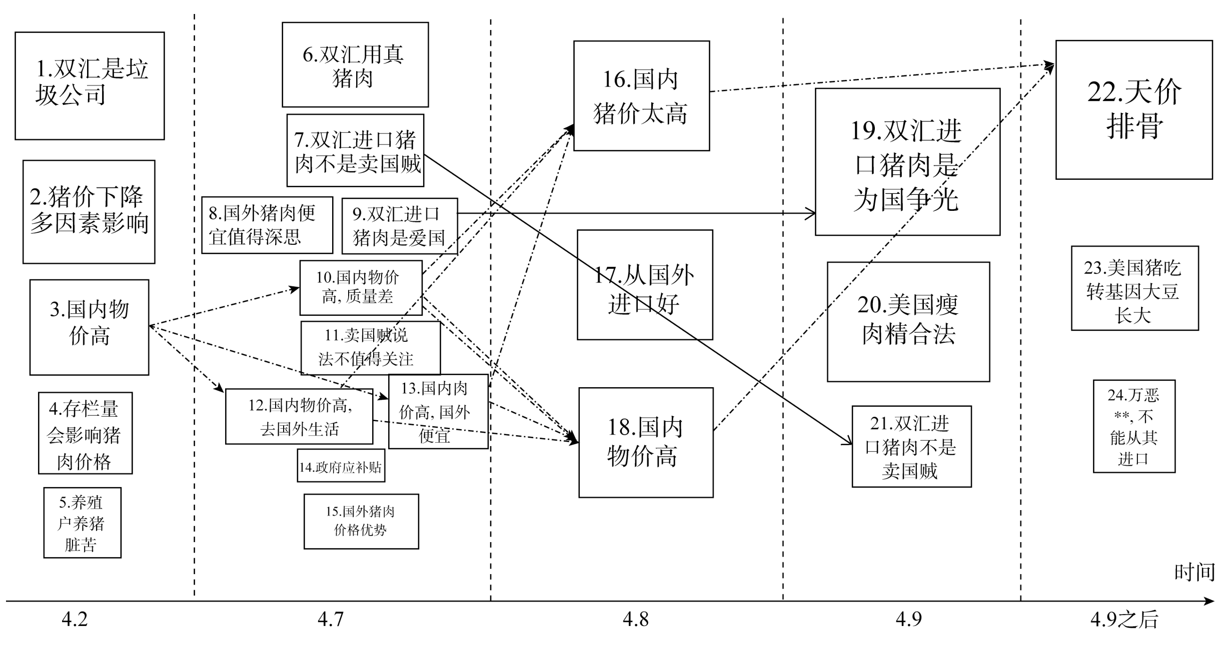


图 13 “双汇进口美国猪肉”事件观点主题演化过程

从图13可以看出，观点主题会随时间的推进而不断变化，但关乎人们自身利益的观点，如人们对物价高的抱怨，会贯穿事件始终。因此当这种关乎公众利益的观点出现，尤其是表现为负面情感倾向时，应该立刻引起政府和企业的重视，及时做好舆情引导工作，避免这种负面观点的继续蔓延。此外，网民观点受其自身认知影响，如，有网民盲目希望所有东西都从国外进口，也有网民能理性地提出美国瘦肉精合法而质疑进口猪肉的质量。当“瘦肉精”相关观点被提出后，又有用户紧接着提出了“美国使用转基因大豆作为猪饲料”这一观点，说明用户观点会受其他用户影响。

5 结 语

本文基于社会网络视角，利用LDA主题模型，多方面、多维度地提出一种网络舆情观点主题识别模型。实验整体效果显示，本文所构建的网络舆情观点主题识别模型能有效识别网络舆情中的观点主题，把握网民主体的主流观点。本研究尚处于初探阶段，结合模型本身及实验效果来看，本文构建的观点主题识别模型还缺少对主题数、词向量个数的确定方法的研究，缺少对主题抽取结果及聚类效果的科学评价。今后，笔者将就上述不足对模型进行不断完善，同时考虑将用户属性引入到观点主题识别的方法体系中，多方位地识别网络舆情观点主题，以期帮助政府和企业了解社情民意，把握网络舆论倾向，做出正确决策。

参考文献：

[1] Wu W, Zhang B, Ostendorf M. Automatic Generation of Personalized Annotation Tags for Twitter Users[C]// Proceedings of the 2010 Annual Conference of the North American Chapter of Association for Computational Linguistics, Los Angeles, California, USA. Association for Computational Linguistics, 2010: 689-692.

[2] Narang K, Nagar S, Mehta S, et al. Discovery and Analysis of Evolving Topical Social Discussions on Unstructured Microblogs[A]// Advances in Information Retrieval [M]. Berlin, Heidelberg: Springer, 2013: 545-556.

[3] Kim H G, Lee S, Kyeong S. Discovering Hot Topics Using Twitter Streaming Data Social Topic Detection and Geographic Clustering[C]//Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis & Mining, Niagara, Ontario, Canada. New York, USA: ACM, 2013: 1215-1220.

[4] Nguyen D T, Jung J E. Privacy-preserving Discovery of

Topic-based Events from Social Sensor Signals: An Experimental Study on Twitter[J]. Scientific World Journal, 2014, 67(3): 435-444.

[5] Guo J, Zhang P, Tan J L, et al. Mining Hot Topics from Twitter Streams[J]. Procedia Computer Science, 2012, 9(11): 2008-2011.

[6] 叶川, 马静. 多媒体微博评论信息的主题发现算法研究[J]. 现代图书情报技术, 2015(11): 51-59. (Ye Chuan, Ma Jing. Topic Discovery Algorithm for Multimedia Microblog Comments Information[J]. New Technology of Library and Information Service, 2015(11): 51-59.)

[7] 唐晓波, 房小可. 基于文本聚类与 LDA 相融合的微博主题检索模型研究[J]. 情报理论与实践, 2013, 36(8): 85-90. (Tang Xiaobo, Fang Xiaoke. Micro Blog Topic Retrieval Model Research Based on Text Clustering and LDA[J]. Information Studies: Theory & Application, 2013, 36(8): 85-90.)

[8] 伍万坤, 吴清烈, 顾锦江. 基于 EM-LDA 综合模型的电商微博热点话题发现[J]. 现代图书情报技术, 2015(11): 33-40. (Wu Wankun, Wu Qinglie, Gu Jinjiang. Research on Hot Topic Discovery of Microblog Based on EM-LDA Comprehensive Model [J]. New Technology of Library and Information Service, 2015(11): 33-40.)

[9] 黄炜, 程宝生, 杨青. 基于本体的网络群体性事件主题发现研究[J]. 图书情报工作, 2012, 56(20): 47-52. (Huang Wei, Cheng Baosheng, Yang Qing. Topic Discovery of Network Group Events Based on Ontology[J]. Library and Information Service, 2012, 56(20): 47-52.)

[10] Huang S, Liu Y, Dang D. Burst Topic Discovery and Trend Tracing Based on Storm[J]. Physica A: Statistical Mechanics and Its Applications, 2014, 416: 331-339.

[11] 夏梦南, 杜永萍, 左本欣. 基于依存分析与特征组合的微博情感分析[J]. 山东大学学报: 理学版, 2014, 49(11): 22-30. (Xia Mengnan, Du Yongping, Zuo Benxin. Micro-blog Opinion Analysis Based on Syntactic Dependency and Feature Combination [J]. Journal of Shandong University: Natural Science, 2014, 49(11): 22-30.)

[12] Deng J, Deng K, Li Y, et al. Hot Topic Detection Based on Complex Networks[C]//Proceedings of the 10th International Conference on Fuzzy Systems and Knowledge Discovery. 2013.

[13] Yin Z, Cao L, Gu Q, et al. Latent Community Topic Analysis: Integration of Community Discovery with Topic Modeling[J]. ACM Transactions on Intelligent Systems and Technology, 2012, 3(4): 67-83.

[14] 王曰芬, 杭伟梁, 丁洁. 微博舆情社会网络关键节点识别与应用研究[J]. 情报资料工作, 2016(3): 6-11. (Wang Yuefen, Hang Weiliang, Ding Jie. Identification and Application of Microblog Public Opinion Social Network Critical Node[J]. Information and Documentation Services, 2016(3): 6-11.)

[15] 周杰, 林琛, 李弼程. 面向网络评论的观点主题识别研究[J]. 情报学报, 2010, 29(5): 858-863. (Zhou Jie, Lin Chen, Li Bicheng. Research on the Identification of Opinion Topic Expressed in Web Comments [J]. Journal of the China Society for Scientific and Technical Information, 2010, 29(5): 858-863.)

[16] 丁晟春, 王颖, 李霄. 基于 SVM 的中文微博情绪分析研究[J]. 情报资料工作, 2016(3): 28-33. (Ding Shengchun, Wang Ying, Li Xiao. SVM-based Chinese Microblog Sentiment Analysis[J]. Information and Documentation Services, 2016(3): 28-33.)

[17] 陈晓美, 高钺, 关心惠. 网络舆情观点提取的 LDA 主题模型方法[J]. 图书情报工作, 2015, 59(21): 21-26. (Chen Xiaomei, Gao Cheng, Guan Xinhui. LDA Theme Model Method for the Extraction of Network Public Opinion[J]. New Technology of Library and Information Service, 2015, 59(21): 21-26.)

[18] 姚兆旭, 马静. 面向微博话题的"主题+观点"词条抽取算法研究[J]. 现代图书情报技术, 2016(7-8): 78-86. (Yao Zhaoxu, Ma Jing. Research on Topic Extraction Algorithm Based on "Topic + Opinion" for Microblog [J]. New Technology of Library and Information Service, 2016(7-8): 78-86.)

作者贡献声明:

丁晟春: 提出研究思路, 设计研究方案;
李真: 具体设计研究, 论文起草;
王楠: 参与设计研究方案, 论文最终版本修订。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: todingding@163.com。
[1] 李真, 丁晟春, 王楠. weibo&comments.xlsx. 观点主题抽取微博及评论语料.

收稿日期: 2017-05-31
收修改稿日期: 2017-07-16

Identifying Topics of Online Public Opinion

Li Zhen Ding Shengchun Wang Nan
(Department of Information Management, Nanjing University of Science and Technology, Nanjing 210094, China)

Abstract: [**Objective**] This paper aims to identify the topics of online public opinion. [**Methods**] We constructed a model to extract public opinion based on the information content of the Weibo posts, the relationship among the users, and user behaviors. [**Results**] We built a public opinion network, extracted and clustered relevant topics, constructed a two-mode network of “user-topic” and evolution of the opinion topics. The proposed method could identify topics of online public opinion effectively. [**Limitations**] The influence of users’ attributes on topic identification needed to be investigated. [**Conclusions**] We could identify the topics of online public opinion based on the social network analysis with the help of LDA model.

Keywords: Network Public Opinion Social Network LDA Model Topic Identification Opinion Topic

学术研究: 机器学习可以预测约会的吸引力程度, 但无法找到完美的灵魂伴侣

约会网站经常声称两个人之间的吸引力大小可以通过爱好和偏好的正确组合来进行预测, 但是一项新的研究却对这一断言提出质疑。该研究分析了快速约会的数据, 发现机器可以预测谁会是你喜欢的, 以及你喜欢她/他的程度, 但是它并不能解释人与人之间的那种狂热喜爱的奥妙所在。

该研究题为《浪漫欲望是否可预测? 机器学习应用于初始浪漫吸引力预测》, 已在 *Psychological Science* 杂志在线发表。研究人员使用两个样本的快速约会数据, 这些受试者填写了 100 多种爱好和偏好的调查问卷, 然后被安排在一系列的 4 分钟约会中见面。之后, 受试者对他们的互动给出评价, 对他们遇到的每个人的感兴趣程度和对他们的吸引力进行打分。

该文章作者心理学教授 Samantha Joel 和其同事使用最先进的机器学习算法测试是否可以根据受试者的问卷回答, 在他们见面之前预测他们是否是彼此的“那个人”。答案是否定的。研究发现预测一个人喜欢和被喜欢的整体趋势是可能的, 但为两个特定的人配对是不可能的。

Joel 说: “我们无法预测在快速约会的环境下能成功匹配多少对。在 100 多个受试者中, 我们原本以为至少可以预测那么几对或十几对, 但是没想到我们的结果竟然是零。”

Joel 认为, 如果人们能够通过将信息输入计算机来寻找完美的灵魂伴侣, 从而克服约会过程中的种种麻烦和心痛, 那将是一件伟大的事情。他表示, 虽然在线约会网站通过缩小范围并识别潜在的恋爱对象来提供有价值的服务, 但是你仍然要通过物理接触这个过程来了解对这些潜在对象的真实感受。

(编译自: https://www.sciencedaily.com/releases/2017/08/170830132200.htm)

(本刊讯)